ALGORITMO CON COBERTURA MUESTRAL EN DATA MINING APLICADO AL ESTUDIO DE LA BIODIVERSIDAD

Cristóbal R. Santa María Departamento de Ingeniería UNLAM
Marcelo Soria Facultad de Agronomía Cátedra de Microbiología UBA
Florencio Varela 1903 San Justo Pcia. de Buenos Aires
54-011-44808952
csantamaria@unlam.edu.ar
soria@agro.uba.ar

RESUMEN

Enmarcadas en la biología computacional, la aplicación conjunta de técnicas de Data Mining y Simulación a secuencias muestrales de ADN con el objeto de evaluar la riqueza, principal parámetro de biodiversidad, ha producido resultados que mejoran las estimaciones usualmente realizadas por procedimientos solo estadísticos. A partir del agrupamiento jerárquico de secuencias de la muestra en diferentes "clusters" que representan taxones distintos seleccionados por umbral de disimilaridad, es posible construir un modelo experimental y aplicar sobre él algoritmos de recuento de especies, o más generalmente de taxones (ARE [1] y [2]), que elevan a niveles compatibles con la apreciación biológica la riqueza subestimada por los procedimientos estándar [3]. Se desarrolla aquí en detalle un algoritmo alternativo a dichos procedimientos ARE que incorpora el concepto de cobertura muestral [4] y proporciona así estabilidad a la simulación asociada. Se procesan dos conjuntos muestrales y se obtienen conclusiones sobre el desempeño del algoritmo con cobertura muestral

Palabras Clave: Cluster-Riqueza-Diversidad-Simulación-Cobertura

CONTEXTO

La línea de investigación que se presenta está inserta en el proyecto Aplicaciones de Data Mining al Estudio de la Biodiversidad en Relevamientos Metagenómicos que, dentro del marco del Programa de Incentivos a la Investigación, se lleva adelante en el Departamento de Ingeniería e Investigaciones Tecnológicas de la UNLAM. Tal tarea se realiza con la colaboración de un investigador de la Cátedra de Microbiología de la Facultad de Agronomía de la UBA y con el asesoramiento de la Maestría en Explotación de Datos y Descubrimiento del Conocimiento de la UBA.

INTRODUCCIÓN

A partir de muestras de material biológico se obtienen secuencias de ADN que son cadenas de símbolos representantes de diferentes componentes químicas y correspondientes a distintos microorganismos. A partir de allí los estudios de biodiversidad devienen en procesos computacionales que abarcan diferentes técnicas. En primer lugar se efectúa un alineamiento y filtrado de esas secuencias con la intención de poder comparar fragmentos de ADN correspondientes a similares lugares del genoma [5]. Es posible realizar estos procesos con software libre disponible en la web tales como MOTHUR [6] Luego se realiza un agrupamiento jerárquico de secuencias , es decir de individuos, de acuerdo a la "distancia" biológica que haya entre ellas. Para evaluar la distancia se recurre a alguno de los distintos modelos disponibles. En este caso se utilizó la distancia del modelo de Jukes-Cantor que posee propiedad ultramétrica. y mide la similaridad entre secuencias [7]. Se empleó el software libre DNADIST de la suite PHYLLIP [8] que calculó la matriz de distancias entre secuencias y con ellas se realizó el agrupamiento

jerárquico según el criterio de encadenamiento promedio. Se utilizó para esto MOTHUR con un nivel de disimilaridad del 5% que permite identificar el taxón especie en el dendrograma correspondiente.

Los distintos grupos así obtenidos corresponden a individuos que por similaridad genética pertenecen a la misma especie. Es decir cada "cluster" representa una especie y habrá en la muestra representada computacionalmente tantas especies como "clusters" se hayan formado.

A partir de aquí el procedimiento estándar [4] para estimar la cantidad de especies en la población de individuos a partir de la muestra hallada produce, a criterio de biólogos y ecólogos, una subestimación de la riqueza medida en términos de cantidad de especies estimada [3]. Esto obedece a las características de la distribución real de las especies microbianas en la población que suele contener una pequeña proporción de especies muy abundantes, otra más pequeña de especies menos frecuentes y una alta cantidad de especies estadísticamente raras, es decir, muy poco frecuentes. Este tipo de especies, que suelen ser la mayoría, complica la estimación de riqueza a partir de muestras, dada su baja probabilidad de aparecer en una de ellas.

Los algoritmos de recuento de especies en los que se trabajó al intentar mejorar la estimación, de lo que se dio cuenta en [1] y [2], lograron alcanzar mejoras significativas en la estimación al tener en cuenta las dos características de la biodiversidad: cantidad de especies y distribución de las mismas.

La idea conductora de tales procedimientos fue el estimador desarrollado por Turing [9] para evaluar la probabilidad de especie nueva al considerar una muestra de tamaño n. Se construyeron así procedimientos de simulación que agregaban individuos a las muestras, los que en algunos casos correspondían a especies simuladas nuevas y aumentaban por ende los valores de riqueza. Esto también proporcionaba una distribución simulada de especies que modelaba la distribución real de la población con las especies raras ahora incluidas. Estos algoritmos produjeron estimaciones sensiblemente mayores que las aportadas por los métodos CHAO y ACE [4] y convergieron a valores más aceptables para el criterio biológico. [1] y [2].

Sin embargo estos nuevos procedimientos, que como es de práctica en data mining investigan a partir de un gran volumen de datos obtenidos, poseen la incertidumbre propia de no poder ser constrastados con los parámetros estadísticos de una población real, pues ello es económica y tecnológicamente imposible habida cuenta de los millones de microorganismos que pueden existir en un determinado medio. Ante esto cabe considerar varias alternativas concurrentes que busquen atenuar la incerteza. Una de ellas es la explorada en el presente trabajo y consiste en tener en cuenta, al hacer crecer simuladamente la muestra, la cobertura supuesta que debe ir alcanzando respecto de la distribución poblacional.

LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

En [4] Chao y Lee presentan la idea de cobertura que aplican para derivar el estimador conocido como ACE. También Chao y Bunge utilizan la misma idea en [10] para construir un estimador del número de especies a partir de modelos de abundancia de tipo paramétrico no pensados, en principio, para comunidades microbianas como las que se analizan.

Cada una de las S especies existentes en el medio tiene una probabilidad p_j de aparición. Si se

toma una muestra de tamaño i, de forma tal que a cada especie le correspondan x_j individuos de la misma, se define la cobertura como $C = \sum_{j=1}^{S} p_j I[x_j \neq 0]$ dónde I es la función indicador que vale 1 si $x_j \neq 0$ y 0 en otro caso. El valor de S es desconocido y, en realidad, en la expresión de C solo suman aquellas especies que efectivamente aparecen. Claramente $0 \leq C \leq 1$ Si $C = 0$ es porque no ha aparecido aún ninguna especie (caso solo teórico e imposible si se tomó una muestra) y si $C = 1$ es porque todas las especies existentes han aparecido en la muestra.

Además a partir de la muestra puede calcularse el número de especies representadas por r individuos

$$f_r = \sum_{j=1}^{S} I(x_j = r)$$

suponiendo una cantidad f_0 que sea precisamente el número de especies con 0 individuos. Obsérvese que puede ocurrir que $f_r = 0$ para varios valores de $r = 1, \ldots, i$ Así $\sum_{t=1}^{y} f_r = S_i$ dónde S_i es la cantidad de especies halladas en la muestra que tiene i individuos y claramente resulta $S = S_i + f_0$ Además si f_1 es el número de singletones en la muestra, es decir el número de especies (clusters) representadas por un solo individuo y $\sum_{r=1}^{i} r f_r = i$ es el tamaño muestral.

Según exponen Chao y Shen [11] un estimador de la cobertura según la muestra tomada es

$$\hat{C} = 1 - \frac{f_1}{i} = 1 - \hat{T}_i$$

La cantidad $\hat{T}_i = \frac{n^{\circ} singletones}{i-1}$ es la estimación de Turing de que si se elige un i-ésimo individuo cuando la muestra tiene tamaño $i-1$, éste sea de una especie nueva [9]. A su vez la probabilidad p_j de elección de un individuo de la j-ésima especie se estima por $\hat{p}_j = \frac{x_j}{i}\hat{C}$

Algoritmo de Recuento de Especies con Cobertura (AREC)

1- Dada la muestra elegida, de tamaño n, y su agrupamiento en especies (clusters), se determina el valor inicial del estimador de Turing $\hat{T}_{i+k} = \frac{f_1}{i}$ siendo $i = n$

2- Se calcula la estimación de cobertura mediante $\hat{C} = 1 - \frac{f_1}{i} = 1 - \hat{T}_{i+1}$

3- Como la muestra actual tiene una cobertura estimada $\hat{C}$ y cada especie que resultó un singletón tiene una frecuencia relativa en la muestra dada por $\frac{1}{i}$, esta probabilidad puede ser corregida por la cobertura estimada de la muestra de modo que la probabilidad de cada singleton resultará $\frac{1}{i}(1 - \frac{f_1}{i})$ Como esto debe ocurrir para los f_1 singletones hallados se obtiene una probabilidad de especie nueva corregida

$$p_{ns} = \hat{T}_{i+1}\hat{C} = \frac{f_1}{i}(1 - \frac{f_1}{i})$$

4- Se elige un número aleatorio r, tal que $0 \leq r \leq 1$ y se pregunta si está en el intervalo $[0, p_{ns}]$ Si es así, se realiza $S_{i+1} = S_i + 1$ y se va al paso 6. Si ocurre lo contrario se realiza $S_{i+1} = S_i$ y se va al paso 5

5- Se utiliza la distribución de abundancia de la muestra, sin corrección por cobertura, para calcular la proporción de individuos que están en especies (clusters) de $1, 2, \ldots, i$ individuos. Con estas proporciones se determina, por un sorteo de acuerdo a ellas, a que grupo de especies (clusters) ya conocidas pertenece el nuevo individuo. Para establecer a que especie (cluster) específica, de entre las de este

grupo, corresponde el nuevo individuo se realiza un nuevo sorteo con probabilidad uniforme para cada especie (cluster) del grupo.

6- Sea el nuevo individuo de una nueva especie o no, la muestra tiene ahora un elemento más. Se pregunta entonces si el procedimiento debe cortarse porque se cumple el criterio elegido para ello, en cuyo caso la simulación ha finalizado. Si el criterio de corte no se cumple, se asigna entonces $i \leftarrow i+1$, se calcula la nueva distribución de abundancia, la nueva estimación de Turing según $\hat{T}_{i+1} = \frac{f_1}{i}$ y se repite desde el paso 2.

El algoritmo fue programado en lenguaje R [12]

RESULTADOS Y OBJETIVOS

Se seleccionaron dos conjuntos de muestras a efecto de probar en ellas el desempeño de las estimaciones de riqueza realizadas por medio de las técnicas disponibles usualmente y de compararlas, luego, con los resultados obtenidos a partir de las ideas y mejoras propuestas.

El primer conjunto corresponde al suelo de La Sal del Rey, región lacustre ubicada en el Estado de Texas, EEUU. Se obtuvieron ocho muestras integradas por secuencias de ADN correspondientes al gen 16S rRNA. Estas secuencias se encuentran almacenadas en NCBI Short Read Archive bajo el número de acceso SRX008158 [13], de donde fueron tomadas para desarrollar el trabajo. El número total de bases químicas almacenadas es de 925673

El segundo conjunto corresponde a suelo de la Selva Amazónica, en Brasil, con tres tipos de manejos. El número de acceso en NCBI Short Read Archive es ERX009564 [12]. Está constituido por seis muestras integradas por secuencias de ADN del gen 16S rRNA cuyo tamaño y nomenclatura se muestra en la Tabla 2. El total del conjunto es de 2400000 bases.

En las Tablas 1 y 2 se presentan los resultados alcanzados por el algoritmo AREC y se listan por muestra, para comparación, las especies observadas (clusters), los valores alcanzados por el procedimiento ARE citado en [2] y las estimaciones estadísticas CHAO y ACE [4]

Tabla 1

Muestra	S85	S86	S87	S88	S89	S90	S91	S92
N° Ind	1641	8361	6926	6146	6226	8444	6103	5885
Clusters	541	3575	2273	2030	1715	2040	1659	1842
CHAO	834	6351	4080	3363	2502	3180	2691	3110
ACE	1102	9525	5964	3755	2755	3544	3757	4048
ARE	1288	9851	6015	5567	3759	4384	3924	4800
AREC	**1088**	**8165**	**5355**	**4743**	**3482**	**4223**	**3586**	**4286**

Tabla 2

Muestra	Er19	Er20	Er21	Er22	Er23	Er24
N° Ind	5011	5582	7637	3299	10371	5840
Clusters	957	1066	1930	997	2664	1358
CHAO	1821	1786	3374	1687	4937	2104
ACE	2239	2458	4531	2125	6237	2129
ARE	2291	2526	5244	2861	7124	3253
AREC	**2175**	**2390**	**4618**	**2452**	**6440**	**3015**

La estimación realizada por AREC es, en todos los casos, levemente inferior a la calculada por ARE pero a pesar del suavizado implícito resulta superior a la mejor estimación estadística (ACE) para más de la mitadde las muestras analizadas. Las Figuras 1 y 2 exhiben esta relación comparativa.

Figura 1

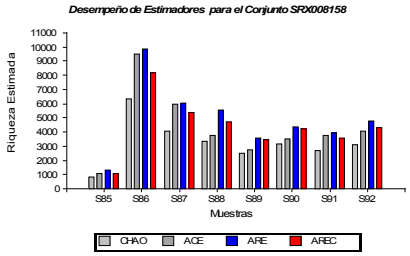
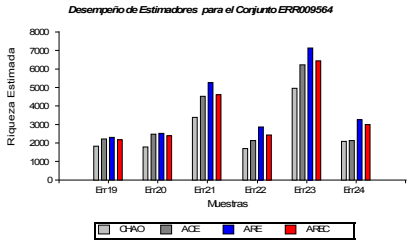


Figura 2



Si bien tanto el procedimiento ARE analizado en [2] como el AREC aquí detallado estiman valores superiores de riqueza compatibles con la expectativa del biólogo y no obstante que ambas construcciones se basan en la estimación de Turing de la probabilidad de especie nueva formalmente probada en [9], cabe señalar que, por las razones ya apuntadas, no se cuenta con población real alguna para el testeo de los resultados. Por esa razón el trabajo se orienta actualmente a construir una población simulada para evaluar estadísticamente con mayor precisión, el desempeño de estos métodos.

FORMACION DE RECURSOS HUMANOS

En 2011, Cristóbal Santa María obtuvo el titulo de Magister en Explotación de Datos y Descubrimiento del Conocimiento que expide la Universidad de Buenos Aires, por el trabajo desarrollado en esta línea de investigación.

REFERENCIAS

1-Santa María, C "Aplicaciones de Data Mining al Estudio de la Biodiversidad en Relevamientos Metagenómicos". Tesis de Maestría. Facultad de Ciencias Exactas y Naturales. UBA. 2011

2-Santa María, C y Soria, M "Estimación de Biodiversidad por Data Mining y Simulación". CACIC. 2011.

3-Roesch, L, Fulthorpe, R, Riva, A, Casella, G, Hadwin, A, Kent, A, Daroub, S, Camargo, F, Farmerie, W y Triplett, E. "Pyrosequencing enumerates and contrasts soil microbial diversity". The ISME Journal. 1, 283-290. 2007

4-Chao, A y Lee, S. "Estimating the Number of Classes via Sample Coverage". Journal of American Statistical Association. Volume 87. Issue 417. 1992

5-Durbin, R, Eddy, S, Krogh, A y Mitchison, G. Biological Sequence Analysis. Cambridge University Press. 1998

6-Schloss, P.D., et al., Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75(23):7537-41. 2009

7- Swofford, D. Olsen, G. Waddell, P. y Hillis, D. Molecular Systematics. Chapter 11. Phylogenetic Inference. Second edition. Edited by David M. Hillis, Craig Moritz, and Barbara K. Mable. 1996

8- http://evolution.genetics.washington.edu/phylip.html

9-Good, I. "The Population Frequencies of Species and Estimation of Population Parameters". Biometrika. Vol 40 N° 3/4. 1953

10-Chao, A. y Bunge, J. "Estimating the Number of Species in a Stochastic Abundance Model". Biometrics 58 Pgs. 531-539. 2002

11- Chao, A y Shen, T. "Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample". Environmental and Ecological Statistics 10, 429-443. 2003

12- http://www.r-project.org/

13- http://www.ncbi.nlm.nih.gov/